1. Based on the data from the file canopycover.txt, the purpose of the study is to examine how the canopy cover Y_i in a forest area i depends on the basal area $x_{i1} =$ basalarea and on the average breast height diameter $x_{i2} =$ dbh.mean of trees in area i, when the main species in the area is either $j = 1 =$ pine or $j = 2 =$ spruce.

```
    species basalarea dbh.mean canopycover
1      pine    20.320   21.797       0.242
2      pine    16.585   15.527       0.148
3      pine     6.708   21.860       0.132
.
114  spruce    16.948   14.482       0.353
```

The variable Y_i measures how many percent of the crown of trees cover the forest area (converted on interval (0,1)).

(a) Choose appropriate distribution and link function $g(\mu_i)$ for modeling the values of random variables Y_i. What is your choice?

i. $Y_i \sim N(\mu_i, \sigma^2)$ and identity link,
ii. $Y_i \sim Gamma(\mu_i, \phi)$ and $\log$ link,
iii. $Y_i \sim IG(\mu_i, \phi)$ and $\log$ link,
iv. $Y_i \sim Beta(\mu_i, \phi)$ and logit link.

(1 point)

(b) Consider modeling the expected value μ_i with the model

$$g(\mu_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \alpha_j,$$

by using your choice of link function g. In the model, parameters α_j are related to the variable $X_3 =$ species. Calculate the maximum likelihood estimate for the expected value μ_{i_*} when $x_{i_*1} = 20, x_{i_*2} = 15$, and $x_{i_*3} =$ pine.

(1 point)

(c) Consider modeling the expected value μ_i with the model

$$g(\mu_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \alpha_j.$$

Calculate the 95% confidence interval estimate for the expected value μ_i when the explanatory variables are set on values $x_{i_*1} = 20, x_{i_*2} = 15$, and $x_{i_*3} =$ pine.

(1 point)

(d) Consider modeling the expected value μ_i with the model

$$g(\mu_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \alpha_j.$$

Create 80% prediction interval for the new observationy_f when the explanatory variables are set on values $x_{f1} = 20, x_{f2} = 15$, and $x_{f3} =$ pine.

(1 point)

(e) Test at 5% significance level, is the explanatory variable $X_2 =$ dbh.mean statistically significant variable in the two way interaction model

$$g(\mu_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \alpha_j + \gamma_j x_{i1} + \delta_j x_{i2}$$

Calculate the value of the test statistic.

(1 point)

2. Based on the data in file NitrogenYield.txt, model how the nitrogen content in fertilization $X =$ Nitrogen affects the amount of yield measured in variable $Y =$ Yield.

```
   Nitrogen  Yield
1        10  28.08
2        10  30.92
3        10  30.71
4        20  37.86
.
59      200 105.55
60      200  92.17

Description: Pounds of Nitrogen (lbs/acre) and yield (bushels) for plots.
X=Nitrogen/acre  (lbs), Y=Yield (bushels)
Source: P.R. Johnson (1953). "Alternative Functions for Analyzing
a Fertilizer-Yield Relationship", Journal of Farm Economics, Vol. 35, #4, pp 519-529.
```

Let us assume the normality $Y_i \sim N(\mu_i, \sigma^2)$.

(a) Model the expected value μ_i by the second degree polynomial model

$$\mu_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2.$$

Calculate the maximum likelihood estimate for the parameter β_2.

(1 point)

(b) Model the expected value μ_i by the exponential model

$$\mu_i = e^{\beta_0} x_i^{\beta_1}.$$

Calculate the maximum likelihood estimate for the expected value μ_{i_*}, when $x_{i_*} = 150$.

(1 point)

(c) Model the expected value μ_i by the asymptotic regression, SSasymp, model

$$\mu_i = \beta_0 + (\beta_1 - \beta_0) e^{(-e^{\beta_2} x_i)},$$

where $\beta_0, \beta_1, \beta_2$ are unknown parameters. Calculate the maximum likelihood estimate for the parameter β_0.

(1 point)

(d) Model the expected value μ_i by the Michaelis-Menten, SSmicmen, model

$$\mu_i = \frac{\beta_1 x_i}{\beta_0 + x_i},$$

where β_0, β_1 are unknown parameters. Calculate the maximum likelihood estimate for the expected value μ_{i_*}, when $x_{i_*} = 150$.

(1 point)

(e) Consider again the asymptotic regression, SSasymp, model

$$\mu_i = \beta_0 + (\beta_1 - \beta_0)e^{(-e^{\beta_2} x_i)}.$$

Create 80% prediction interval for new observation y_f, when $x_f = 150$.

(2 points)

3. Consider the data set in the file caffeine.txt:

```
> data<-read.table("caffeine.txt", sep="\t", header=TRUE, dec=".")
> head(data)
  Brand Formulation   Caffeine
1  Coke       Sugar     47.32
2  Coke       Sugar     43.78
3  Coke       Sugar     48.12
4  Coke       Sugar     43.25
.

source: A.N. Garand and L.N. Bell (1997). "Caffeine Content of Fountain and
Private-Label Store Brand Carbonated Beverages," Journal of the
American Dietetic Association, Vol. 97, #2, pp. 179-182.

Description: Caffeine content (mg/12oz) for 2 formulations (sugar/diet)
of 2 Brands (Coca-Cola/Pepsi)

Variables/Columns:
Brand    1=Coke, 2=Pepsi
Formulation   1=Sugar,  2=Diet
```

Denote variables as following

$$Y = \text{Caffeine},\ X_1 = \text{Brand},\ X_2 = \text{Formulation}.$$

(a) Consider modeling the expected value of the response variable $Y = \text{Caffeine}$ by the explanatory variables X_1, X_2. Use methods and models considered during the course. Select the appropriate default distribution for the response variable Y, and consider several competing models. Choose the model which you feel is the most suitable one for modeling the expected value of the response variable $Y = \text{Caffeine}$. The best model you choose does not have to include all possible explanatory variables. In your solution, try to report your modeling process, that is, which models you considered, which link functions you compared, which goodness of fit measures you used, and which hypotheses you tested before you chose your final model. Reporting can mainly be R-code with some clarifying comments. Try make sure that your R-code is running without errors before returning your solution.

(3 points)

(b) After you have chosen your model, construct 80% prediction interval for the new observation Y_f when explanatory variables are set to the values $x_{1f1} =$ Coke, $x_{1f2} =$ Diet.

(2 points)

(c) Test at 5% significance level, is the explanatory variable $X_1 =$ Brand statistically significant variable. In your solution, report clearly how you calculated the value of the test statistic.

(1 point)